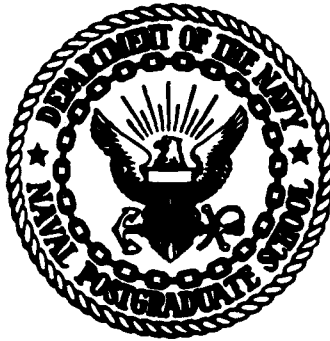


NPS55MT71011A

AD735127

# United States Naval Postgraduate School



DDC  
RECEIVED  
JAN 18 1972  
R  
C

A COMPARISON OF TWO PERSONNEL PREDICTION MODELS

by

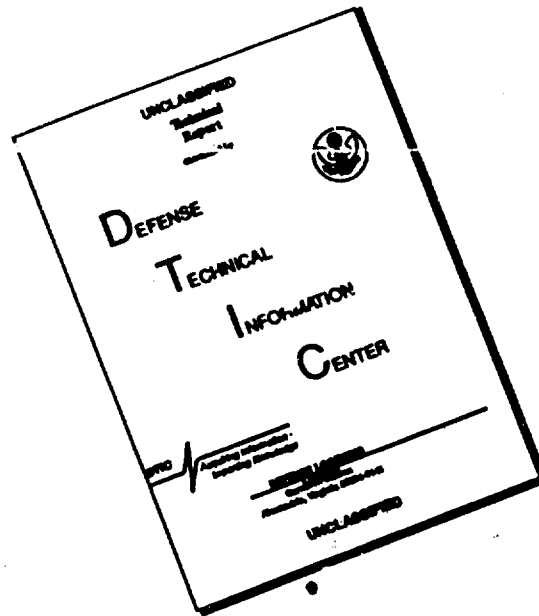
Kneale T. Marshall

January 1971

This document has been approved for public release and  
sale; its distribution is unlimited.

Reproduced by  
NATIONAL TECHNICAL  
INFORMATION SERVICE  
Springfield, Va 22151

# DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

UNCLASSIFIED  
Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) Naval Postgraduate School Monterey, California		2a. REPORT SECURITY CLASSIFICATION Unclassified	
		2b. GROUP	
3. REPORT TITLE A Comparison of Two Personnel Prediction Models			
4. DESCRIPTIVE NOTES (Type of report and inclusive dates) Technical Report 1971			
5. AUTHOR(S) (First name, middle initial, last name) Marshall, Kneale T.			
6. REPORT DATE January 1971		7a. TOTAL NO. OF PAGES 43	7b. NO. OF REFS
8a. CONTRACT OR GRANT NO. NR047-096		8b. ORIGINATOR'S REPORT NUMBER(S) NPS55MT71011A	
9. PROJECT NO.		9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)	
10. DISTRIBUTION STATEMENT This document has been approved for public release and sale; its distribution is unlimited.			
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY	
13. ABSTRACT <p>This paper describes, compares and contrasts two mathematical models of personnel movement through a hierarchical organization. The first model is a Markov chain type which is described in detail in other literature. The emphasis in this paper is on a cohort model based on peoples lifetime behavior in the system. Data from student enrollments is used in comparing the models, and predictions are made and compared with actual numbers.</p>			

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Manpower						
Personnel						
Prediction						
Markov Chains						
Cohort Models						
Retention						

NAVAL POSTGRADUATE SCHOOL  
Monterey, California

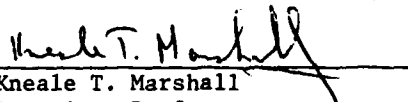
Rear Admiral R. W. McNitt, USN  
Superintendent

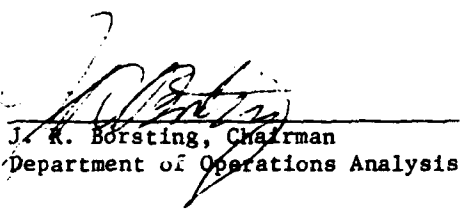
M. U. Clauser  
Provost


ABSTRACT:

This paper describes, compares and contrasts two mathematical models of personnel movement through a hierarchical organization. The first model is a Markov chain type which is described in detail in other literature. The emphasis in this paper is on a cohort model based on people's lifetime behavior in the system. Data from student enrollments is used in comparing the models, and predictions are made and compared with actual numbers.

This research was supported in part by the Office of Naval Research under contract task number NR 047-096, by the Office of Institutional Research, University of California, Berkeley, and the Ford Foundation.

  
Kneale T. Marshall  
Associate Professor

  
J. R. Borsting, Chairman  
Department of Operations Analysis

  
C. E. Menneken  
Dean of Research Administration

NPS55MT71011A

January 1971

## TABLE OF CONTENTS

	Page
I Introduction	1
II The Markov Model	3
III A Cohort Model	5
IV Model Comparison	11
V Enrollment Forecasts	22
Appendix	25
References	27

## I. Introduction.

The purpose of this paper is to describe, compare and contrast two mathematical models used to describe movement of personnel through a hierarchical organization.

The first model, which has received considerable attention in the literature (for example, see Bartholomew (1967), Gani (1963), Thonstad (1968)) assumes an underlying stationary Markov chain structure. The important point of this type of model is that it uses crosssectional data of an organization in a given time period, and predicts what will be the composition of the organization (i.e., the cross section) in the following time period(s). A major advantage of such a method is that it requires little data.

The second model considered here is of the cohort type. This model follows each group of newly entering people, called a cohort, over their lifetimes in the organization. Cross sectional structure in any time period is found by considering the super-position of the remaining members of all the previously entering cohorts. Although more appealing from a theoretical viewpoint, this model typically requires considerably more data than the Markov model.

The Markov model is described briefly in section II and the Cohort model in detail in section III. In section IV an attempt is made to compare theoretically the two models. Under certain conditions the models give essentially the same results. The results of the analysis show that under stationary conditions the Markov method gives a good approximation to the movement through an organization, and since its data requirements are small such a model may be preferred. However,

for organizations with changing or controlled cohort sizes the fractions which appear in the Markov method should be changed from year to year and the model gives no functional relationship of the model parameters to the sizes of the cohorts. In the cohort method the parameters appear as functions of the cohort sizes, and so in non-stationary situations, the cohort method may be preferred for long range forecasting.

In section V some enrollment predictions are made, using both the Markov chain and Cohort Models, of student enrollments at the University of California, Berkeley. These are compared with actual enrollments.



## II. The Markov Model.

The Markov chain model has been discussed in detail in the literature (for example, see Bartholomew (1967), Gani (1963), Marshall et al (1970)), but to unify notation, and for completeness and clarity we formulate it here. Throughout the paper we assume a system made up of  $n$  active states.

At each time period it is assumed that people can stay in the same grade, can move to other grades, or can leave the system. New inputs are added to the continuing or promoted people in each grade. Possible movement is shown schematically in figure 1 for 3 grades.

Let  $X_i(t)$  be the number in grade  $i$  at time  $t$ ,  $i \in P$ , and let  $X(t)$  be a row vector  $(X_1(t), \dots, X_n(t))$ , where  $|P| = n$ . Let  $E_m[X(t)]$  be the vector of expected numbers in each rank at  $t + 1$ , given the vector  $X(t)$ . Then

$$E_m[X(t)] = X(t)Q(t) + y(t+1), \quad (1)$$

where

$$Q(t) = \begin{bmatrix} q_{11}(t), q_{12}(t), \dots, q_{1n}(t) \\ q_{21}(t), \\ \vdots \\ q_{n1}(t), q_{n2}(t), \dots, q_{nn}(t) \end{bmatrix}, \quad (2)$$

and

$$y(t) = (y_1(t), \dots, y_n(t)).$$

The vector  $y(t+1)$  is a vector of new inputs into each grade at time period  $(t+1)$ , i.e.,  $y_i(t+1)$  = number new people who enter grade  $i$  at  $t + 1$ . The matrix  $Q(t)$  has the structure of the transient part of a Markov chain matrix, and  $q_{ij}(t)$  is the fraction of those in  $i$  at  $t$  who will move to  $j$  at  $t + 1$ .

The main advantage of this model is that only a small amount of data is required to estimate the coefficients; only the grade of each person in the last two time periods is required.

Although the name "Markov-Chain" method gives the connotation of a stochastic model, in most instances this model is treated in the literature in terms of expected values only, and hence can be considered to be deterministic. However, using the probabilistic interpretation of the Markov chain, it is assumed that the probability a person is promoted to state  $j$ , given he is now in  $i$ , is independent of how long he has been in  $i$ , or how he got into  $i$ . This seems an unreasonable assumption. In section III we formulate a "cohort" model of movement through a system of grades which is based on more reasonable assumptions. In section IV we compare the two models.

$$Q = \begin{bmatrix} .5 & .4 & 0 \\ 0 & .6 & .3 \\ 0 & 0 & .7 \end{bmatrix}.$$

All new input at  $t + 1$  into grade 1,  $y_1(t+1) > 0$ ,  $y_2(t) = y_3(t) = 0$ .

All new input at  $t + 2$  into grade 2,  $y_2(t+2) > 0$ ,  $y_1(t) = y_3(t) = 0$ .

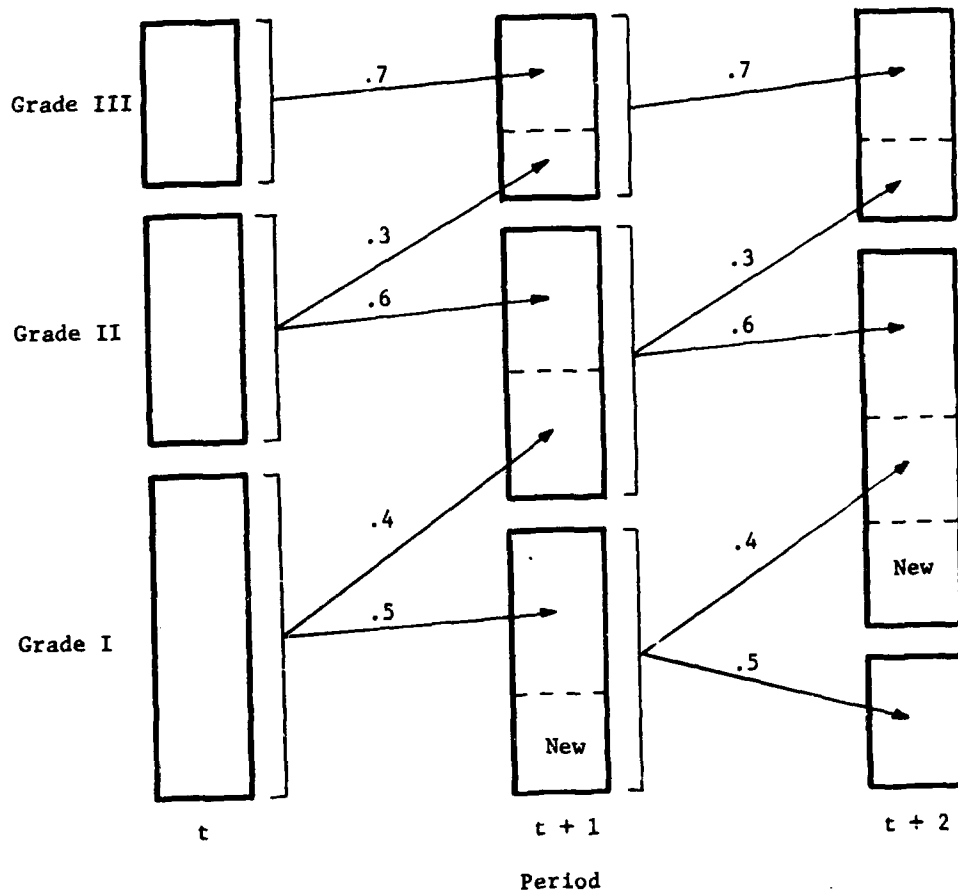


Figure 1: Illustration of Markov Chain Model with 3 grades.

### III. A Cohort Model.

People who enter into a system in the same grade and in the same time period are referred to as a cohort. For example, all freshmen entering a given university in a particular academic quarter, or all officers entering into the U.S. Navy as Ensigns with regular commission in a given fiscal year would be considered in each case to form a cohort.

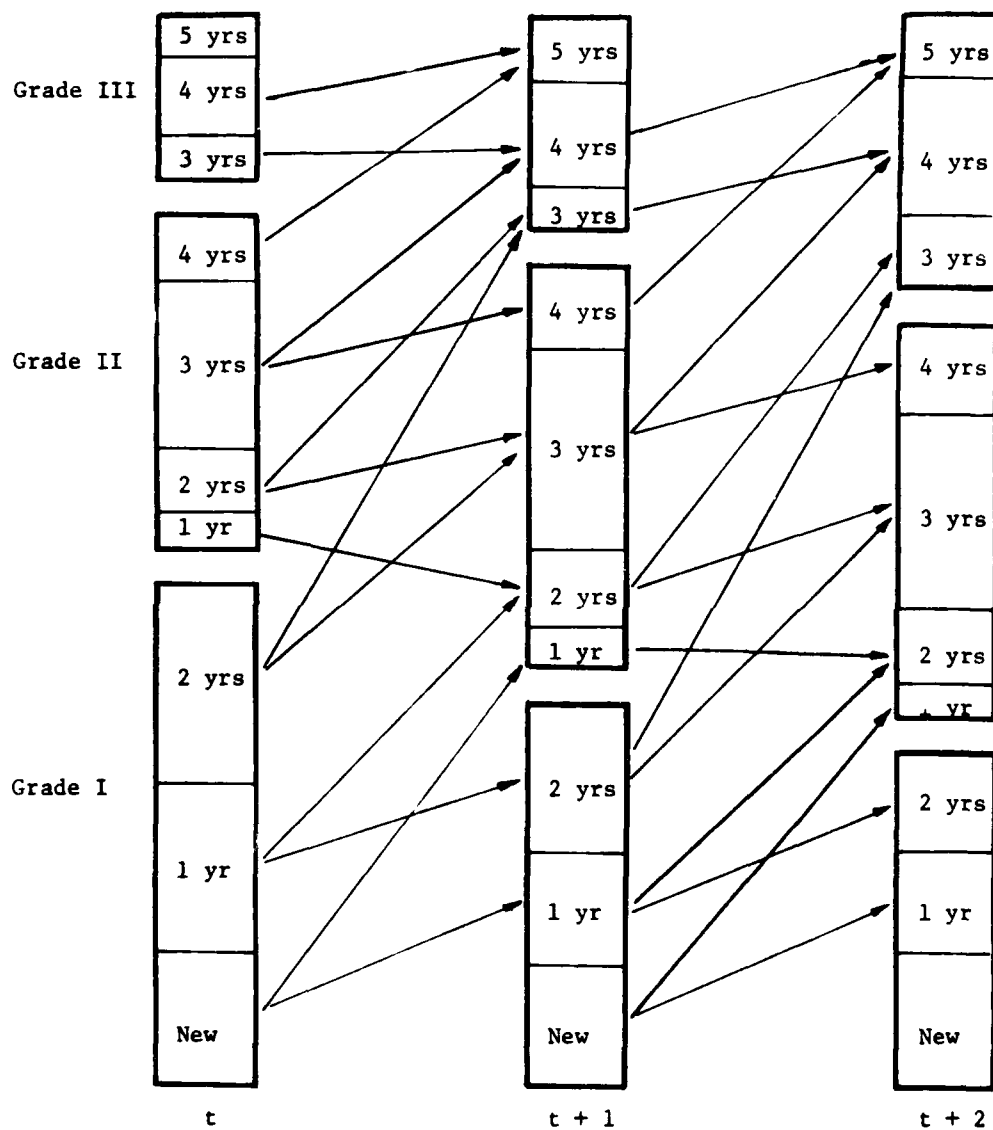
After some time the people in a given cohort will be found in various grades in the system, and some will have left. We can think of the people in a given grade at some time as coming from many previously entering cohorts. Indeed, everyone in the system entered in some cohort. The cross-sectional structure in a given time period can be thought of as the result of the superposition of the remnants of all previously entering cohorts. Figure 2 gives a schematic representation of the cohort model.

Let there be  $n$  different types of cohorts which enter the system. For example, students can enter a university as freshmen, sophomores, juniors, or seniors. Let  $y_i(u)$  be the number who enter in cohort  $i$  at time  $u$ . Let  $k$  index the people in a given cohort. Thus define

$$z_{ij}^{(k)}(u,t) = \begin{cases} 1 & \text{if person } k \text{ of cohort } i \text{ which} \\ & \text{entered at } u \text{ is in } j \text{ at } t, \\ 0 & \text{otherwise,} \end{cases}$$

for  $k = 1, 2, \dots, y_i(u)$ .

We shall assume that all cohorts behave independently of each other and that all members of a given cohort have independent behavior. Let  $z_i^{(k)}(u,t) = (z_{i1}^{(k)}(u,t), \dots, z_{in}^{(k)}(u,t))$ ,  $k = 1, 2, \dots, y_i(u)$ . Thus



**Figure 2:** Illustration of Cohort Model with 3 grades.

we have a set of  $y_i(u)$  independent and identically distributed  $n$ -dimensional vectors.

Now let

$$p_{ij}(u,t) = \Pr[Z_{ij}^{(k)}(u,t) = 1], \quad (3)$$

and

$$P(u,t) = [p_{ij}(u,t)],$$

the  $n \times n$  matrix. Then

$$E[Z_i^{(k)}(u,t)] = (p_{i1}(u,t), \dots, p_{in}(u,t)).$$

Since we are interested in relating the positions of people in consecutive time periods, define the  $2n$ -vector

$$\begin{aligned} & [Z_i^{(k)}(u,t), Z_i^{(k)}(u,t+1)] \\ &= [Z_{i1}^{(k)}(u,t), \dots, Z_{in}^{(k)}(u,t), Z_{i1}^{(k)}(u,t+1), \dots, Z_{in}^{(k)}(u,t+1)]. \end{aligned}$$

Let  $X_{ij}(u,t)$  be the number of people in  $j$  at  $t$  who entered in  $i$  at  $u$ . Also let  $[X_i(u,t), X_i(u,t+1)]$  be the  $2n$ -vector of  $X_{ij}(u,t)$ ,  $X_{ij}(u,t+1)$ ,  $j = 1, 2, \dots, n$ . Then

$$[X_i(u,t), X_i(u,t+1)] = \sum_{k=1}^{y_i(u)} [Z_i^{(k)}(u,t), Z_i^{(k)}(u,t+1)]. \quad (4)$$

From our assumptions this vector is the sum of  $y_i(u)$  independent and identically distributed vectors, and thus for large cohort sizes the  $[X_i(u,t), X_i(u,t+1)]$ ,  $i = 1, 2, \dots, n$ ,  $u < t$ , are each approximately normally distributed (see for example, chapter 4 of Anderson (1958)). We shall assume that cohorts are large enough for normality assumptions to hold.

Let  $X_j(t)$  be the number in grade  $j$  at time  $t$  and let  $X(t) = (X_1(t), \dots, X_n(t))$ . Then

$$[X(t), X(t+1)] = \sum_{u \leq t} \sum_{i=1}^n [X_i(u, t), X_i(u, t+1)] + [0, y(t+1)], \quad (5)$$

where  $y(t+1)$  is the  $n$ -vector of new inputs at  $t+1$ , and  $0$  is an  $n$ -vector of zeros. Again we have a sum of independent random vectors. They are not identically distributed, but if each is approximately normal, then the  $2n$ -vector  $[X(t), X(t+1)]$  has a multivariate normal distribution. In terms of the original  $Z$  vector random variables,

$$[X(t), X(t+1)] = \sum_{u \leq t} \sum_{i=1}^n y_i(u) \sum_{k=1}^n [Z_i^{(k)}(u, t), Z_i^{(k)}(u, t+1)] + [0, y(t+1)]. \quad (6)$$

In forecasting, what we need is the conditional expectation  $E[X(t+1)|X(t)]$ . It is well known that (see Anderson (1958), chapter 2) for the multivariate normal distribution,

$$E[X(t+1)|X(t)] = E[X(t+1)] + [X(t) - E[X(t)]]B(t)^{-1}C(t), \quad (7)$$

where  $B(t)$  is the  $n \times n$  covariance matrix of elements of  $X(t)$ , and  $C(t)$  is the  $n \times n$  covariance matrix of elements of  $X(t)$  with corresponding elements of  $X(t+1)$ .

To compare this result with equation (1) we let  $E[X(t+1)|X(t)] = E_c[X(t)]$ , and write (7) as

$$\begin{aligned} E_c[X(t)] &= X(t)B^{-1}(t)C(t) \\ &+ y(t+1) + [E[X(t+1)] - y(t+1) - E[X(t)]B^{-1}(t)C(t)]. \end{aligned} \quad (8)$$

Equation (8) has the same linear structure as (1), but the coefficients appear to be quite different from those of the Markov chain model. We explore this further in section IV. However, we shall need to know the structure of  $B(t)$  and  $C(t)$  in more detail, and now find them in terms of the cohort sizes and the underlying probability distributions.

### Structure of $B(t)$ .

Recall that  $B(t)$  is the covariance matrix of the elements of  $X(t)$ . Thus  $b_{ij}(t) = \text{Cov}[X_i(t), X_j(t)]$  where  $X_i(t)$  is the number in state  $i$  at time  $t$ . From (6) we have

$$\text{Cov}[X(t), X(t+1)] = \sum_{u \leq t} \sum_{i=1}^n y_i(u) \sum_{k=1}^n \text{Cov}[(Z_i^{(k)}(u, t), Z_i^{(k)}(u, t+1))].$$

The expression for  $B(t)$  in terms of the original probability distributions is given in equation (9). Note that  $B(t)$  is symmetric with off diagonal terms negative and diagonal terms positive. Now define  $\mu_j(t) = E[X_j(t)]$ , the expected number in state  $j$  at time  $t$ . Then  $\mu_j(t) = \sum_{u \leq t} \sum_{i=1}^n y_i(u) p_{ij}(u, t)$ . Let  $M(t)$  be the diagonal matrix with diagonal elements  $\mu_i(t)$ . Also define  $Y(u)$  to be an  $n \times n$  diagonal matrix with diagonal elements  $y_i(u)$ . With these definitions (9) simplifies considerably and we have

$$B(t) = M(t) - \sum_{u \leq t} P(u, t)^T Y(u) P(u, t), \quad (10)$$

where  $T$  denotes transpose and the  $P$  matrices are given by (3).



$$B(t) = \sum_{u \leq t} \sum_{i=1}^n y_i(u)$$

$$\begin{bmatrix} [1-p_{i1}(u,t)]p_{i1}(u,t), & -p_{i1}(u,t)p_{i2}(u,t), & \dots, & -p_{i1}(u,t)p_{in}(u,t) \\ -p_{i2}(u,t)p_{i1}(u,t), & [1-p_{i2}(u,t)]p_{i2}(u,t), & \dots, & -p_{i2}(u,t)p_{in}(u,t) \\ \vdots & \vdots & \ddots & \vdots \\ -p_{in}(u,t)p_{i1}(u,t), & -p_{in}(u,t)p_{i2}(u,t), & \dots, & [1-p_{in}(u,t)]p_{in}(u,t) \end{bmatrix}$$

Equation (9).

Structure of  $C(t)$ .

Earlier we defined  $C(t)$  to be the covariance matrix of elements of  $X(t)$  with those of  $X(t+1)$ . Thus  $c_{j\ell}(t) = \text{Cov}[X_j(t), X_\ell(t+1)]$ .

Define the joint distribution

$$\pi_{ij\ell}(u, t) = P[Z_{ij}^{(k)}(u, t) = 1, Z_{i\ell}^{(k)}(u, t+1) = 1],$$

all  $k = 1, \dots, y_i(u)$ . Then

$$c_{j\ell}(t) = \sum_{u \leq t} \sum_{i=1}^n y_i(u) [\pi_{ij\ell}(u, t) - p_{ij}(u, t)p_{i\ell}(u, t+1)]. \quad (11)$$

Let  $\lambda_{j\ell}(t)$  be the expected number of people who move from grade  $j$  at  $t$  to grade  $\ell$  at  $t+1$ , and let  $\Lambda(t) = [\lambda_{j\ell}(t)]$ , an  $n \times n$  matrix. Then from (11) and the definition of  $\Lambda$

$$C(t) = \Lambda(t) - \sum_{u \leq t} P(u, t)^T Y(u) P(u, t+1). \quad (12)$$

#### IV. Model Comparison.

In this section we compare the two estimators  $E_m$  and  $E_c$  for the Markov and Cohort models respectively. Taking the stochastic interpretation of the Markov model we see that

$$q_{j\lambda}(t) = \lambda_{j\lambda}(t)/\mu_j(t). \quad (13)$$

Thus from (1) we have

$$E_m = X(t) \cdot M^{-1}(t) A(t) + y(t+1). \quad (14)$$

For the cohort model, from (7) we have

$$E_c = X(t) B^{-1}(t) C(t) + y(t+1) + [\mu(t+1) - y(t+1) - \mu(t) B^{-1}(t) C(t)]. \quad (15)$$

Now  $\mu_j(t+1) = \sum_{i=1}^n \lambda_{ij}(t) + y_j(t+1)$ , and assume that we can pick  $y_j(t+1)$  so that  $\mu_j(t+1) = \mu_j(t)$  for all  $j$ . Then

$$\mu_j(t) = \sum_{i=1}^n \lambda_{ij}(t) \frac{\lambda_{ij}(t)}{\mu_i(t)} + y_j(t+1),$$

or

$$\mu(t) - y(t+1) = \mu(t) Q(t). \quad (16)$$

Using (16) together with (14) and (15) we find that

$$E_m - E_c = [\mu(t) - X(t)] [B^{-1}(t) C(t) - Q(t)]. \quad (17)$$

Equation (17) is useful in comparing the two models. If in some period  $t$  the actual distribution of personnel coincides with the expected distribution, the models will give the same forecasts for period  $t+1$ . "On average" the difference between the two models' forecasts will be zero but for a given period the difference will depend on the size of  $[B^{-1}(t) C(t) - Q(t)]$ .

Using (12) we can write

$$C(t) = A(t) - F(t), \quad (18)$$

where we have let

$$F(t) = \sum_{u \leq t} P(u, t)^T Y(u) P(u, t+1).$$

Similarly, from (10)

$$B(t) = M(t) - G(t), \quad (19)$$

where

$$G(t) = \sum_{u \leq t} P(u, t)^T Y(u) P(u, t).$$

Using (18) and (19) with (13) we find that

$$\{B^{-1}(t)C(t) - Q(t)\} = B^{-1}(t)[G(t)Q(t) - F(t)]. \quad (20)$$

Now if motion through the system is Markovian (possibly non-stationary), then

$$P(u, t+1) = P(u, t)Q(t),$$

and the expression in (20) is zero. This shows the expected result that if motion through a graded system is truly Markovian then the cohort model and Markov chain model give identical forecasts.

Since movement between grades is typically non-Markovian, we wish to investigate further the error given by (17). We shall do this by looking further at  $G(t)Q(t) - F(t)$  for some special cases.

Single Grade Case.

Let us consider the case where we have:

- A1. The system has a single grade ( $n=1$ ),
- A2. At each time period all input cohorts are the same ( $y(t)=y$ ),
- A3. The life distribution of each person in the system is stationary.

With these assumptions the models and their corresponding notation simplify considerably. No subscripts are required on the distribution  $p$ , and if  $L(u)$  is the lifetime in the system of a person entering at  $u$ , then

$$\begin{aligned}\Pr[L(u) > t - u] &= p(u, t) \\ &= p(t-u) \text{ under A3.}\end{aligned}$$

If  $y$  is the constant cohort size for  $u \leq t$  (we cannot claim  $y(t+1) = y$  and that (17) holds simultaneously), then

$$\begin{aligned}G &= \sum_{u \leq t} yp(t-u)^2, \quad M = \sum_{u \leq t} yp(t-u). \\ \Lambda &= \sum_{u \leq t} yp(t+1-u), \quad F(t) = \sum_{u \leq t} yp(t-u)p(t+1-u).\end{aligned}$$

All these are independent of  $t$ .

Now let  $\ell = E[L] = \sum_{u \leq t} p(t-u)$ . Then

$$GQ - F = \frac{Y}{\ell} \left[ \sum_{u=0}^{\infty} p(u)^2 \sum_{u=0}^{\infty} p(u+1) - \sum_{u=0}^{\infty} p(u+1)p(u) \sum_{u=0}^{\infty} p(u) \right]. \quad (21)$$

The term in parenthesis in (21) is

$$\sum_{u \geq 0} p(u)^2(\ell-1) - \ell \sum_{u \geq 0} p(u)p(u+1) = \ell \sum_{u \geq 0} \Delta(u+1)p(u) - \sum_{u \geq 0} p(u)^2, \quad (22)$$

where  $\Delta(u+1) = P[L = u + 1] = p(u) - p(u+1)$ .

Interpreting  $p(u)$  as the tail distribution of a non-negative random variable, one can show that

$$\sum_{u \geq 0} p(u)[1 - p(u)] = \sum_{u \geq 0} \Delta(u) \sum_{v \geq u} p(v), \quad (23)$$

and

$$\sum_{u \geq 0} [\Delta(u) + \Delta(u+1)]p(u) = 1. \quad (24)$$

Using (22), (23) and (24) in (21) gives

$$GQ - F = \frac{\gamma}{\ell} \sum_{u \geq 0} \Delta(u) \left[ \sum_{v \geq u} p(v) - (\ell)p(u) \right]. \quad (25)$$

Let us assume now that the expected remaining lifetime of a person whose time in the system exceeds  $u$  time periods is no more than the expected lifetime  $\ell$  of a new input. We say that people have "mean residual life" bounded above by the original mean life, and say that  $L$  has MRLA if

$$\sum_{v \geq u} \frac{p(v)}{p(u)} \leq \ell, \quad \text{all } u = 0, 1, 2, \dots \text{ for which } p(u) > 0.$$

Note that equality holds in this equation for the geometric distribution.

Table 1 shows that in a particular case of students attending the University of California at Berkeley, this assumption is valid.

Under the MRLA assumption, from (25) we see that

$$GQ - F \leq 0. \quad (26)$$

Recall that

$$E_m - E_c = [\mu - X(t)]B^{-1}[GQ - F]. \quad (27)$$

Since  $B^{-1}$  is non-negative, we have the following conclusion under the above four assumptions:

If in addition to A1 - A3 we assume  $L$  has MRLA,

- a) If  $X(t) < \mu$ , then  $E_m \leq E_c$  and the Markov model under-estimates the value of  $E[X(t+1)|X(t)]$ ,
- b) If  $X(t) > \mu$ , then  $E_m \geq E_c$ , and the Markov model over-estimates the value of  $E[X(t+1)|X(t)]$ .

TABLE 1: Mean Residual Life of Freshmen Students Entering  
U.C. Berkeley in Fall Semester, 1955.

Lifetime (semesters) $u$	$P_r[L > u]^*$ $= p(u)$	$\sum_{v \geq u} p(u)$	$\sum_{v \geq u} p(u)/p(v)$
0	1.000	6.959	6.96
1	0.972	5.959	6.14
2	0.905	4.987	5.52
3	0.756	4.082	5.42
4	0.684	3.326	4.86
5	0.593	2.642	4.47
6	0.562	2.049	3.65
7	0.524	1.487	2.84
8	0.498	.936	1.88
9	0.199	.465	2.34
10	0.130	.266	2.05
11	0.050	.136	2.72
12	0.036	.086	2.39
13	0.017	.050	2.94
14	0.015	.033	2.20
15	0.011	.018	1.64
16	0.007	.007	1.00

\* Source data found in Suslow et al (1968), [5].

Since  $X(t)$  has a marginal normal distribution we can say more about the expected error in the one dimensional case.  $(E_m - E_c)$  is a normal random variable with zero mean, and variance equal to  $B^{-1}(GQ-F)^2$

(where these are all scalars). Thus we can say that with probability about .95 the error  $(E_m - E_c)$  will lie in the interval  $(-2 B^{-1/2} |GQ - F|, +2 B^{-1/2} |GQ - F|)$ . The length of this interval is a function of the cohort size  $y$ , and increases as  $y^{1/2}$ . The expected value of  $X(t)$ ,  $\mu$ , increases as  $y$ . Thus the interval length divided by  $\mu$ , or the fractional error range, decreases as  $y^{1/2}$ . So as  $y$  increases, and hence  $\mu$  increases, the width of the confidence interval of error increases much more slowly. To illustrate this we use the lifetime distribution from table 1, and for various cohort sizes we show how the interval length changes. The results are given in table 2. It is clear from this table that even though the lifetime distribution differs considerably from a Markovian (geometric) distribution with the same mean, the confidence intervals on  $E_m - E_c$  are extremely small relative to the expected number in system,  $\mu$ . For comparison  $p(u)$  is drawn in figure 3 together with a geometric distribution.

TABLE 2: 95% Confidence Intervals for  $E_m - E_c$   
for various Cohort Sizes.

Cohort Size $y$	$E[X]$ $= \mu$	Confidence* Interval for $E_m - E_c$
1000	6,959	(-7,7)
2000	13,918	(-10,10)
3000	20,877	(-12,12)
4000	27,836	(-14,14)

\*Based on lifetime distribution in table 1.



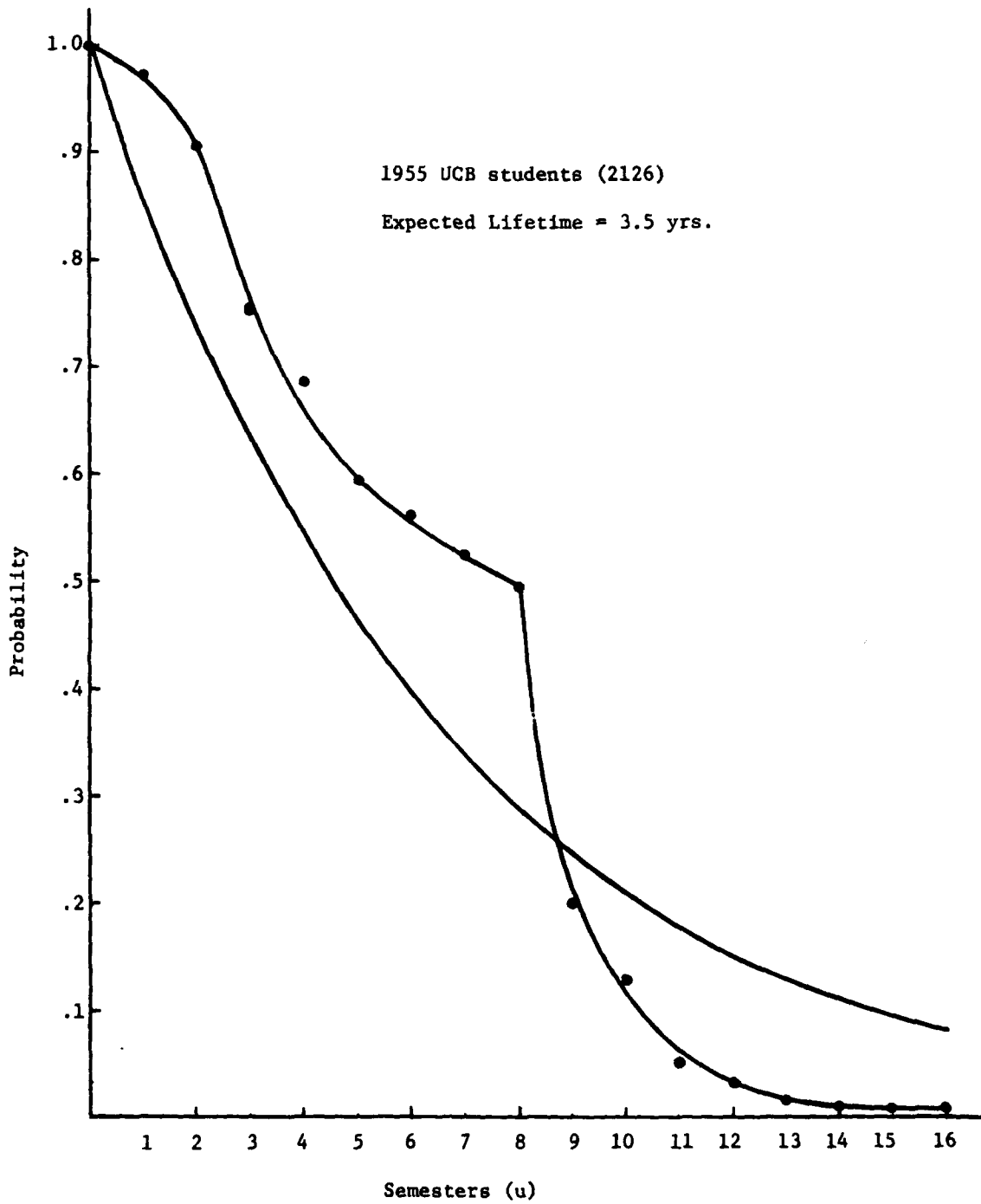


Figure 3: Comparison of  $p(u)$  for UCB Students with a geometric distribution.

Multigrade System.

Let us now relax assumption A1, but keep the assumptions A2 and A3 of constant cohort sizes and stationary distributions respectively. Let  $Y$  be the diagonal matrix of cohort sizes at each time period. Define  $L = \sum_{u \geq 0} P(u)$ , where  $P(t-u) = P(u, t)$ . Under such stationary conditions  $\mu(t) = \mu$  independent of  $t$ , and if  $y$  is the  $n$ -vector of cohort sizes, then from expected value arguments

$$\mu Q = \mu - y.$$

Thus

$$\mu = y(I-Q)^{-1} \text{ and also}$$

$$\begin{aligned} \mu &= y \sum_{u \geq 0} P(u) \\ &= y L. \end{aligned}$$

Since these relationships hold for all  $y$ ,  $L = I - Q^{-1}$ , and finally

$$Q = I - L^{-1}. \quad (27)$$

Using (27) with the definitions of  $G$  and  $F$ , we have that

$$GQ - F = \sum_{u \geq 0} P(u)^T Y [P(u)(I - L^{-1}) - P(u+1)]. \quad (28)$$

Recall from (17) and (20) that

$$E_m - E_c = [\mu - X(t)] B^{-1} [GQ - F].$$

It is easy to show that  $B^{-1}$  is non-negative, but the conditions under which  $E_m > E_c$ , or conditions for this inequality to hold for some

element  $i$  are much more complex than in the single state case. Let  $\Delta(u+1) = P(u) - P(u+1)$ . Then the multi-dimensional equivalents of (23) and (24) are

$$\sum_{u \geq 0} [I - P(u)^T] Y(u) P(u) = \sum_{u \geq 0} \Delta(u)^T \sum_{v \geq u} Y(v) P(v), \quad (29)$$

and

$$\sum_{u \geq 0} [P(u)^T Y \Delta(u+1) + \Delta(u)^T Y P(u)] = Y. \quad (30)$$

Note that (30) only holds for  $Y$  a stationary matrix, whereas in (29)  $Y(u)$  can change over time.

Using (29) and (30) in (28) gives as the multidimensional equivalent of (25),

$$GQ - F = \sum_{u \geq 0} \Delta(u)^T Y \left[ \sum_{v \geq u} P(v) L^{-1} - P(u) \right]. \quad (31)$$

Although this equation has great similarity to (25) it is quite different. Even if one can say something about the sign of  $\sum_{v \geq u} P(v) L^{-1} - P(u)$ , it is usually true that  $\Delta(u)$  is not non-negative, as in the single dimensional case. Also of course the elements of  $[\mu - X(t)]$  can differ in sign, so that the conditions for each element of  $E_m - E_c$  to be either negative or positive do not seem simple or natural.

Equation (28) seems to be the most useful for computation purposes. Note that  $(E_m - E_c)$  has a multivariate normal distribution with mean 0 and covariance matrix  $(GQ - F)^T (B^{-1})^T (GQ - F)$ . Using the data given in the appendix for freshmen, sophomores, juniors and seniors at the University of California, Berkeley 1955-1969, some calculations were made assuming constant cohort sizes of 3000 freshmen, 700 sophomores, 1300 juniors

and 150 seniors entering each fall semester. These figures are approximately what the Berkeley campus has been experiencing in its fall new admissions.

Table 3 gives the matrix  $B$ , whose  $(i,j)^{th}$  element is the covariance of  $X_i(t)$  and  $X_j(t)$  for some  $t$ . Also included is  $\mu$ , the vector of expected values of numbers in each state.

TABLE 3: Covariance Matrix  $B$  for the 4-state example.

State i \ State j				
	Fresh	Soph	Jun	Sen
Fresh	673	-454	-30	-10
Soph	-454	1453	-380	-43
Jun	-30	-380	2137	-535
Sen	-10	-43	-535	2216
Expected Values	3868	3324	4687	3227

The variance of the number in each state increases as the state increases, and all states are negatively correlated.

Table 4 gives the matrix  $(GQ-F)^T B^{-1} (GQ-F)$ , which is the covariance matrix of the error  $(E_m - E_c)$ . It can be seen that these numbers are very small compared to the size of the predicted values, as was found in the single state case.

TABLE 4: Covariance Matrix of  $E_m - E_c$ .

State j \ State i		Fresh	Soph	Jun	Sen
Fresh	6.7	2.2	-22.4	-5.4	
Soph	2.2	1.0	-8.5	-2.7	
Jun	-22.4	-8.5	82.2	29.5	
Sen	-5.4	-2.7	29.5	41.8	

The matrix  $B^{-1}(GQ-F)$  is given in table 5.

TABLE 5:  $B^{-1}(GQ-F)$  for the 4-state example.

State j \ State i		Fresh	Soph	Jun	Sen
Fresh		.068	-.041	0.290	.040
Soph		.033	-.003	-.062	-.046
Jun		.002	.003	-.030	-.125
Sen		.001	.001	.029	.032

This is an example of where  $(GQ-F)$  is neither  $\geq$  nor  $\leq 0$ , unlike the single state case.

Even though movement through the system is far from that represented by a stationary Markov Chain, (i.e.,  $P(u) \neq P^u$  for some  $P$ ), when constant cohort sizes are used the Markov Chain Model gives essentially the same prediction as the more complex cohort model.

However, the Cohort Model was primarily formulated for forecasting under conditions of controlled input. This is the situation when academic planning is implemented, and under such conditions the sizes of cohorts in successive time periods can and do vary considerably. For example, the freshmen cohorts in the fall quarters at Berkeley in the period 1966-1969 are shown in table 6. This was a period when total campus enrollment was controlled, and new students entered only to fill available room.

TABLE 6: Freshmen Cohort Sizes at U.C. Berkeley

Date	Cohort Size
Fall 1966	3,053
Fall 1967	3,300
Fall 1968	2,239
Fall 1969	1,883

One can see from equation (13), since  $\lambda(t)$  and  $\mu(t)$  are both functions of previous cohort sizes (up to period  $t$ ), that the Markov chain transition probabilities will change with time, and that estimating them from cross-sectional data in two consecutive years will not account for changes in cohort sizes. In the next section we make forecasts one year ahead with both models and compare the results.

## V Enrollment Forecasts.

In this section we use data up to the spring quarter of 1970 at Berkeley to forecast continuing and returning undergraduate students at the freshman, sophomore, junior and senior levels, in the fall quarter of 1970. Both the Cohort and Markov Chain models are used, and results compared with actual enrollments.

In applying the Cohort Model directly, three problems appeared, all associated with the start-up and operation of the quarter system at Berkeley.

The first winter and summer quarters were offered in 1967. The fractions of students who entered in these quarters and were enrolled in F69 (this notation will be used in this section. F69 means fall quarter 1969) are now applied to cohorts entering in the winter and summer of 1968 when forecasting for F70. It would certainly be expected that some students from the winter and summer quarters of 1967 would also be enrolled in F70, but how many? We have no fractions for winter or summer 1966. These fractions have to be estimated in some reasonable way. An average was taken of the fractions from F65 and Sp66, for the winter quarter and from Sp66 and F66 for the summer quarter.

The third problem that arose was in deciding what fractions to apply to the students who entered in Su69. These students had available only the winter and spring quarters of 1970 before F70. The students who entered in Su68 could attend winter, spring and summer quarters before F69. It was felt that larger fractions of Su69 entrants would attend the fall of 1970 than the fractions of Su68 students attending F69. But how much larger?

To estimate attendance of Su69 entrants it was assumed that the same fraction of these would attend F69 as did Su68 entrants in F68. Of these that enrolled in F69, they were then assumed to behave in the same way as new entrants in F69.

Besides these three particular and rather confusing problems, the stationarity of most of the fractions since the start of the summer quarter can be questioned. With such a major change in campus operations it will take a number of years to settle down even if there were no changes between 3-quarter and 4-quarter operations.

The Markov Chain Model was used in the following way. The transition matrix from F68-F69 was determined by finding the fractions of those enrolled in each grade in F68 who were enrolled in each grade in F69. This matrix is shown in table 7.

TABLE 7: Markov Chain Matrix for F68-F69 at Berkeley

	F69			
	Fr.	So.	Ju.	Se.
Fr.	.162	.551	.066	.001
F68 So.		.105	.640	.035
Ju.			.178	.481
Se.				.152

If this is applied to F69 enrollments, the prediction for F70 will have ignored new inputs in W70 and Sp70 (the summer quarter 1970 was not held). To make a fair comparison the same fractions of these were assumed to enroll in F70 as was assumed in the Cohort Model.



Table 8 shows the forecasts from the two models together with the actual figures. It can be seen that the cohort model gave significantly better predictions than the Markov Chain method. This is not surprising, since these forecasts are made for a period of much instability on the Berkeley campus, both in student behavior and in academic policy.

TABLE 8: Enrollment forecasts for Fall 1970 at Berkeley,  
Continuing and Returning Students

	Freshman	Sophomore	Junior	Senior	Total
Markov Chain Model	958	2,737	4,356	4,189	12,240
Cohort Model	1,115	3,018	4,508	4,670	13,311
Actual	1,591	3,136	4,632	4,261	13,620

APPENDIX

Data used in calculations in table 3. The time periods  $u$  are in years. The data is from many different cohorts, and each number is the fraction of a particular cohort who were enrolled at U. C. Berkeley in the given class in the fall quarter of 1969. Let:

State 1: Freshmen

State 2: Sophomores

State 3: Juniors

State 4: Seniors.

Example:

$p_{13}^{(3)}$  = fraction of students who entered as freshmen in Fall 1966 who registered as juniors in Fall 1969. (0.281).

Time $u$	$P(u)$	$y(u)^T$
0	$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$	1883 258 817 48
1	$\begin{bmatrix} .254 & .584 & .009 & \\ & .118 & .622 & .039 \\ & & .265 & .493 \\ & & & .395 \end{bmatrix}$	2239 542 1366 124

Time	u	P(u)				$y(u)^T$
2	[	.012	.210	.454	.009]	3303
			.013	.189	.337]	843
				.138	.192]	1662
					.046]	175
3	[	.007	.027	.281	.318]	3053
			.003	.022	.130]	733
				.003	.042]	1418
					.029]	205
4	[	.004	.008	.033	.152]	2579
			.003	.005	.031]	390
				.005	.008]	1042
					.016]	125
5	[	.003	.003	.009	.031]	3427
				.003	.010]	602
				.001	.003]	1442
					.015]	202
6	[	.003	.003	.004	.015]	3620
			.001	.004	.007]	728
				.001	.003]	1569
					0]	199

All numbers are rounded off to 3 figures. For more detail see Marshall and Suslow (1971).

REFERENCES

- [1] Anderson, T. W., (1958), *An Introduction to Multivariate Statistical Analysis*, J. Wiley, New York.
- [2] Bartholomew, D. J., (1967), *Stochastic Models for Social Processes*, J. Wiley, New York.
- [3] Gani, J., (1963), "Formulae for Projecting Enrollments and Degrees Awarded in Universities," *J. Roy. Stat. Soc. A.*, 126, pages 400-409.
- [4] Marshall, K. T., Oliver, R. M., Suslow, S., (1970), "Undergraduate Enrollments and Attendance Patterns," Report No. 4, Administrative Studies Project in Higher Education, Office of Institutional Research, University of California, Berkeley.
- [5] Suslow, S., Langlois, E., Sumariwalla, R., Walther, C., (1968), "Student Performance and Attrition at the University of California, Berkeley," Office of Institutional Research, Berkeley.
- [6] Thonstad, T., (1968), *Education and Manpower*, University of Toronto Press.